STUDENT EVALUATION OF TEACHING AND COURSES

TEACHING AND COURSE EVALUATION PROJECT
FINAL REPORT

APPENDIX III: REPORT ON KEY RESEARCH FINDINGS - INSTRUMENTATION

This research report was prepared by Hui Niu as part of the SFU Teaching and Course Evaluation Project. The report has been reformatted for inclusion in the project's final report.

MAY 29, 2012

Contact: Corinne Pitre-Hayes
604.765.4610
cpitreha@sfu.ca
# TABLE OF CONTENTS

1 EXECUTIVE SUMMARY ......................................................................................................................... 1
  1.1 CONTEXT .............................................................................................................................................. 1
  1.2 KEY ISSUES REGARDING INSTRUMENTATION .................................................................................. 1
    1.2.1 MULTIDIMENSIONALITY OF TEACHING EFFECTIVENESS .................................................. 1
    1.2.2 PURPOSES OF EVALUATING TEACHING EFFECTIVENESS .................................................... 1
    1.2.3 VALIDITY AND RELIABILITY ...................................................................................................... 1
    1.2.4 POTENTIAL BIASES: MYTHS OR REALITY? ............................................................................ 2

2 INTRODUCTION ........................................................................................................................................ 3
  2.1 THE PROJECT ..................................................................................................................................... 3
  2.2 THE FIELD ......................................................................................................................................... 3

3 MEASURING TEACHING EFFECTIVENESS ............................................................................................ 5
  3.1 WHAT TO MEASURE? ....................................................................................................................... 5
  3.2 WHY MEASURE? .............................................................................................................................. 5

4 HOW TO MEASURE: INSTRUMENT DEVELOPMENT .............................................................................. 7
  4.1 ITEM CONSTRUCTION ....................................................................................................................... 7
  4.2 QUESTIONNAIRE CONTENT AND FORMAT .................................................................................... 7
  4.3 PSYCHOMETRIC TESTING ............................................................................................................... 8
    4.3.1 VALIDITY ..................................................................................................................................... 9
    4.3.2 RELIABILITY .............................................................................................................................. 10
  4.4 INSTRUMENT IMPLEMENTATION .................................................................................................... 11
  4.5 USE AND INTERPRETATION OF STUDENT EVALUATION DATA .................................................. 11

5 POTENTIAL BIASES: MYTHS OR REALITY? ......................................................................................... 13
  5.1 PRIOR SUBJECT INTEREST .............................................................................................................. 13
  5.2 EXPECTED GRADE/GIVEN GRADES ............................................................................................... 13
  5.3 WORKLOAD/DIFFICULTY .................................................................................................................. 14
  5.4 CLASS SIZE ..................................................................................................................................... 14

6 REFERENCES ............................................................................................................................................. 15
1 EXECUTIVE SUMMARY

1.1 CONTEXT

Student evaluations of teaching and courses have been widely used internationally for various purposes such as tenure and promotion and teaching improvement. The Teaching and Course Evaluation Project (the Project) at Simon Fraser University (SFU) is an initiative sponsored by the VPA on the recommendation of the Senate Committee on University Teaching and Learning (SCUTL) and the Task Force on Teaching and Learning (TFTL) to develop new instruments and guidelines for student evaluation of teaching and courses. This report serves as one of the supporting documents for the Project to provide the SFU community with an overview of research findings and current practices on student evaluation of teaching and courses with a focus on instrument development and related methodological issues. The Project strives to develop a valid and reliable instrument that best suits SFU’s evaluation needs by closely examining the relevant research literature and through community engagement during the process of instrument development and implementation.

1.2 KEY ISSUES REGARDING INSTRUMENTATION

1.2.1 MULTIDIMENSIONALITY OF TEACHING EFFECTIVENESS

Student evaluations of teaching and courses intend to measure a hypothetical construct of “teaching effectiveness.” There is a vast amount of research on student evaluations of teaching, however no consensus has been reached on what constitutes “good teaching.” It is generally understood that teaching effectiveness is a multidimensional and complex construct that includes multiple and possibly networked sub-constructs. What constitutes effective teaching may also vary from institution to institution, discipline to discipline, and faculty to faculty. Therefore, special attention will be given to gathering information on local preferences, priorities and standards pertinent to effective teaching at SFU. Such standards and priorities will be developed based on the literature and SFU’s unique identity and evaluation needs at institutional, departmental, or unit levels. This rationale has prompted the planning of high level community engagement as an essential part of the Project in seeking advice, consultation, and feedback from all stakeholders before and during the process of developing a sound instrument that can best meet SFU’s evaluation needs.

1.2.2 PURPOSES OF EVALUATING TEACHING EFFECTIVENESS

The purpose of evaluating teaching effectiveness is twofold: collecting information for summative and formative reasons. Formative evaluation in the context of teaching quality is intended for instructors to improve their teaching. Summative evaluation is commonly used for personnel decisions on tenure, promotion, or teaching appointments. A larger context for the use of teaching assessment is as a response to the growing interest of the public to hold postsecondary institutions accountable for student learning.

1.2.3 VALIDITY AND RELIABILITY

Numerous research studies, reviews and meta-analyses have shown that student evaluation of teaching can be valid, reliable and stable. However, faculty opinions of student evaluation of teaching vary vastly from viewing it as reliable, valid and useful to unreliable, invalid and useless.

Validity of student evaluation of teaching is primarily dependent upon the interpretation and use of the results, including inferences about teaching effectiveness. Validity has been the most common concern.

---

1 For a broader review on general issues regarding student evaluations of teaching and courses, see the report submitted to the Senate Committee on University Teaching and Learning by Johanne Provençal in January 2012.
2 See the report submitted to the Senate Committee on University Teaching and Learning by Johanne Provençal in January 2012 for more details about this topic.
about student evaluations of teaching in higher education. Such concerns are connected to assumptions that students are not qualified to evaluate teaching and the existence of potential bias variables. These variables have been researched extensively in the literature, and overall, researchers argue that student ratings are consistent with other data sources such as instructor self-evaluation, trained observer evaluation and former student ratings. Although more research is needed to establish validity of student evaluations, such as using more criteria for teaching effectiveness rather than mainly student achievement, researchers are optimistic that student ratings of teaching effectiveness are at least one measure that reflects teaching quality.

Reliability is an important aspect of validity. Research has shown that a student evaluation instrument can be reliable, especially when the instrument is well designed and tested according to psychometric principles. Psychometrics is the field of study concerned with the theory and technique of psychological measurement. The field is primarily concerned with the construction and validation of measurement instruments and includes the development of theoretical approaches to measurement. Reliability of a student evaluation tool is usually determined by inter-rater agreement (agreement across different students in the same class) and internal consistency of a scale (correlations among items measuring the same element of effective teaching)\(^3\). Stability refers to reliability over time. Longitudinal studies comparing immediate ratings with ratings at least a year after have shown high correlations and thus demonstrate good stability.

Generalizability refers to reliability across situations. It assumes that evaluation of an instructor's general teaching effectiveness across courses is more reliable than one snapshot of his/her performance in a particular course at a particular time. Research has indicated that student ratings of teaching effectiveness across different courses or the same course across multiple terms are consistent. It has been recommended that for summative purposes, average scores of several courses taught by the same instructor should be used to enhance generalizability.

1.2.4 POTENTIAL BIASES: MYTHS OR REALITY?

Perceptions of potential biases in student evaluation of teaching can be a source of anxiety for instructors. Important concerns regarding background variables that may threaten validity of student ratings have been discussed extensively in the literature over decades. Some researchers believe these variables introduce biases into the evaluation outcomes, others argue that they are merely sources of validity evidence because they cannot be disconnected from features and dimensions of teaching effectiveness\(^4\). While caution is required in interpreting effects of background variables on student ratings, there seems to be a trend in the literature indicating that some of these biases are not warranted by research evidence. However, it is important for the Project team to take into account evidence-based potential biases when developing and validating our instruments.

\(^3\) For example the scale of an instructor's organization skills consists of four items in SEEQ, a well-researched instrument developed by Herbert Marsh: 1) Instructor's explanations are clear; 2) Course materials are well prepared and carefully explained; 3) Proposed objectives agree with those actually taught so you know where the course is going; 4) Instructor gives lectures that facilitate taking notes.

\(^4\) For example, the effects of expected grades/grading leniency is one of the most common concerns among faculty. However a student's expected grade from a course is likely to correlate with his/her efforts toward and achievement on the subject.
2 INTRODUCTION

2.1 THE PROJECT

The Teaching and Course Evaluation Project is an initiative sponsored by the Vice-President Academic (VPA) on the recommendation of the Senate Committee on University Teaching and Learning (SCUTL) and the Task Force on Teaching and Learning (TFTL) to develop new instruments and guidelines for student evaluation of teaching and courses. The purpose of this document is to provide the Simon Fraser University (SFU) community with an overview of research findings and current practices on student evaluation of teaching and courses with a focus on instrument development and related methodological issues. In order to suit the needs of a wide variety of potential readers (faculty, administrators and students), this report will remain brief and avoid excessive use of technical jargon. For those who are interested in the scientific process applied in the instrumentation exploration and development, a final report documenting the entire instrument development process will be available upon conclusion of the Project.

2.2 THE FIELD

Over the past few decades, much attention has been paid to student evaluation of teaching and courses in higher education. Debates, discussions, and thousands of research articles have been published on the topic. There has been a general understanding in the field of teaching performance in postsecondary institutions that student evaluation is merely one source of evidence to gather on teaching excellence. Other means such as peer and expert observations, teaching philosophy, and teaching portfolios also bring valuable insights into an understanding of instructor's teaching contributions. Nonetheless, in the past 30-40 years, student evaluation of teaching and courses has been the most predominant form of instructional evaluation in higher education throughout the world. Student evaluations are widely used by instructors to improve their teaching (formative) as well as by administrators to make personnel decisions such as merit, tenure and promotions (summative). Recently, more efforts have been made in complementing student ratings with other sources of evidence to better delineate the quality of teaching (such as in Arreola, 2000; Braskamp & Ory, 1994; Knapper & Cranton, 2001).

Although researchers and experts still debate over whether student evaluation is a valid and useful measurement of teaching effectiveness, many empirical studies are generating evidence on valid and
reliable instruments that can be used by students (Sprooner, et al, 2007). In practice, institutions are paying more and more attention to improving and refining their evaluation instruments in compliance with more recent teaching and learning theories as part of their move toward accountability and quality assurance of their institutions\textsuperscript{11}.

\textsuperscript{11} See Report 1 for a fuller discussion on accountability and quality assurance of postsecondary institutions.
3 MEASURING TEACHING EFFECTIVENESS

3.1 WHAT TO MEASURE?

From a measurement point of view (Wilson, 2005), the central goal of constructing a student evaluation instrument is to provide a reasonable and consistent way to summarize student responses that express their attitudes or personal viewpoints on “teaching effectiveness.” Though no consensus has been reached on what precisely constitutes “good teaching”, there has been a general understanding that it is a multidimensional, complex construct that includes multiple and possibly networked sub-constructs. There have been numerous models available from the literature with different numbers of dimensions that represent teaching effectiveness (for example see Kulik & McKeachiem 1975; Feldman, 1976 and Marsh, 1987). Although there are overlapping factors across these models, no common set of dimensionalities has been agreed upon by researchers in the field. Allin, Clarke, and Jopling (2009) attempted to summarize the literature, including the widely cited studies such as Marsh (1987), and generated a four-dimensional structure of effective teaching with twenty-three sub-dimensions (p364)¹²:

- **Supportive Learning Environment**, including sub-constructs such as provision of intellectual excitement; high degree of subject knowledge; respect for students; approachability; provision of a motivating environment; recognition of student diversity, etc.
- **Academic Expectations**, including high level of expected output; explaining to students what they are to learn and why; clarity in standards and assessment criteria; appropriate workload and level of difficulty, etc.
- **Scaffolding Learning**, including varied ways to teach content; anticipation of misconceptions in students’ existing knowledge; appropriate pace for the group being taught; excellent management of student behavior; systematic, well organized and well-structured sessions; students work collaboratively with both their peers and their teachers; effective and timely feedback; effective & sympathetic guidance, etc.
- **Clarity**, including strong, unambiguous presentation skills; high quality explanation.

For the purpose of this project, such models will be contemplated in order to gain insights for our instrument development process. However, as emphasized by researchers there has been no universal set of characteristics of effective teaching that applies to all teaching situations (see Ory & Ryan, 2001 for example), effective teaching may vary from institution to institution, discipline to discipline (Rando, 2001, Cashin, 1990; Feldman, 1978), and faculty to faculty (Ghedin & Aquario, 2008). Special attention will be given to gathering information on local preferences, priorities and standards pertinent to effective teaching at SFU. Such standards and priorities will be developed based on the literature and SFU’s unique identity and evaluation needs at institutional, departmental or unit levels. This rationale has prompted the planning of high level community engagement¹³ as an essential part of the project in seeking advice, consultation, and feedback from all stakeholders before and during the process of developing a sound instrument that can best meet SFU’s evaluation needs.

3.2 WHY MEASURE?

Scriven (1991) defines evaluation as “the process, whose duty is the systematic and objective determination of merit, worth, or value. Without such a process, there is no way to distinguish the worthwhile from the worthless” (p. 4). The purpose of evaluating teaching effectiveness is twofold: collecting information for summative and formative purposes. Formative evaluation in the context of teaching quality is intended for instructors to improve their teaching. Research has shown that student

¹² This model is cited here merely as an illustration of the multidimensionality and complexity of teaching effectiveness. The Project is not endorsing superiority of this particular model over any other ones in the literature.

¹³ See another document on community engagement authored by Chris Groeneboer (Manager, Applied Research, Teaching and Learning Centre, SFU).
feedback from evaluation of teaching and courses can help improve quality of instruction (Cohen, 1980; Marsh & Roche, 1993; Menges, 1991; Overall & Marsh, 1979) and consequently student learning (Goldschmid, 1978). It is noteworthy that such benefits do not come automatically without additional efforts, such as open discussions with colleagues and administrators about student feedback (Penny & Coe, 2004; Spooren, et al, 1997). Faculty may also benefit more from student evaluations if they are provided with training or assistance on data interpretation.

Summative evaluation is used for decisions on “merit, worth, or value”. In this context it means personnel decisions on tenure, promotion, or teaching appointments. Research has shown that student evaluations can provide valuable information on rewarding teaching excellence (Aleamoni, 1981; McKeachie, 1979). However it has been generally understood that student evaluation should not be taken as the only indicator of effective teaching when making personnel decisions. There are many other sources of information to gather on teaching excellence which should be considered in triangulation with student ratings.

A larger context for the use of teaching assessment is as a response to the growing interest of the public to hold postsecondary institutions accountable for student learning. Student ratings have been found to be positively associated with student learning and achievement (Aleamoni & Hexner, 1980; Centra, 1976, 1977, 1987; Cohen, 1981; McKeachie, 1990; Murray, et al, 1990). Although critics argue that some correlations have only been moderate (Derry, 1979; McCallum, 1984) or inconsistent among studies (Dowell & Neal, 1982), there is cause for optimism. For a fuller discussion on such matters, see a separate report. 14

14 See Report 1 prepared by Johanne Provencal.
4 HOW TO MEASURE: INSTRUMENT DEVELOPMENT

4.1 ITEM CONSTRUCTION

Researchers assert that, item composition and selection is very important in ensuring usefulness when measuring teaching effectiveness (Marsh, 2007). However, in practice, item construction is sometimes not carefully executed when institutions develop their instruments locally. “Poorly worded or inappropriate items will not provide useful information” (Marsh & Roche, 1997, p1187). Researchers advise that evaluation questions should avoid the following pitfalls: double-barreled questions (asking two questions at the same time, such as “Rate the assignments and the exams.”), overly complex or ambiguous wording (such as “rate the quality of this course”), or poorly scaled response options (Franklin, 2001). Student evaluations usually include Likert type of scales (e.g., from 1-5), quality rating scales (e.g., fair, good, excellent), and frequency rating scales (e.g., how often does the instructor...?). It is important that students understand the meaning of the scales as intended. For example on a 5-point Likert scale, students may have different interpretations of the middle category 3 and thus affect evaluation outcomes (Ory & Ryan, 2001).

4.2 QUESTIONNAIRE CONTENT AND FORMAT

The content and format of instruments are important factors that determine the evaluation outcomes. Instrument development should be firmly grounded on educational and learning theories, practice and research which have been evolving over time. As emphasized by Theall and Franklin (2000), pedagogical changes in postsecondary education call for updates of evaluation tools that were developed when lecturing was the main form of delivery and students were of homogenous backgrounds. For example, changes in instructional practices include collaborative learning, change in student demographics and online programs, and change in evaluation data collection methods. Items that put too much weight on traditional teaching methods may under represent the characteristics of effective teaching in an innovative, active learning environment that employs methods such as collaboration, discussion-oriented instruction as opposed to lectures, and the use of computers for instruction (Ory & Ryan, 2001).

According to the literature, it is desirable to include items on multiple aspects of teaching behaviors due to the multidimensional nature of teaching effectiveness. Multidimensionality implies that an instructor may be proficient with some components of good teaching but not with others (Algozzine et al., 2004; Centra, 1993; Marsh, 1987; Marsh & Roche, 1997), for example, high on clarity but low on enthusiasm; therefore a sound instrument needs to address a range of fundamental elements of successful teaching with representative questions. Although there has been no consensus on what these fundamental elements should be, researchers have preferences based on their epistemologies of teaching and learning, and universities develop their own standards based on their historical contexts. For example, Kansas State University developed Individual Development and Educational Assessment (IDEA) with rating forms for teacher behaviors that include course organization and planning, clarity/communication skills, teacher student interaction/rapport, course difficulty/workload, grading and examinations and student self-rated learning (Braskamp and Ory,1994). Murray (1987) developed the Teaching Behaviors Inventory, which groups teaching performance into 9 categories: clarity, expression, interaction, organization, pacing, disclosure, speech, rapport, and teaching aids. The well-researched Students' Evaluation of Educational Quality (SEEQ) instrument (Marsh, 1984, 1987; Marsh & Dunkin, 1992) is based on nine factors of effective teaching: learning/value, instructor, enthusiasm, organization/clarity, group interaction, individual rapport, breadth of coverage, examinations/grading, assignments/readings, and workload/difficulty.

In practice, institutions use different formats in structuring and delivering student evaluations. However, there are some common features:

(a) A combination of open-ended and closed-ended questions. Closed-ended questions are designed to collect quantitative responses that can be used for norms, distributions and comparisons. Open-ended
questions are designed to give students the opportunity to further elaborate on their experiences. Rando (2001) offered an example of an open-ended question, entitled ‘the one-minute paper’: “What’s the most important thing you learned and what question are you left with?” Such questions allow more freedom on student responses and can generate feedback on student personal experiences. Such feedback is usually kept confidential to the instructor for formative purposes. Researchers tend to disagree on whether qualitative data should be used for summative purposes. For example, Abrami (2001) asserted that they should not be used for personnel decisions because it is difficult to estimate validity and reliability of qualitative data, while Harper and Kuh (2007) argued otherwise. It is noteworthy that research has shown a correlation between quantitative and qualitative outcomes (Cashin, 1995).

(b) At least one “global” item addressing ‘overall’ effectiveness (such as “overall I learned a great deal from this instructor”) with specific items on multiple perspectives on teaching and course content.

For example, at McGill, student evaluation includes 4 global questions:

- Overall, this is an excellent course.
- Overall, I learned a great deal from this course.
- Overall, this instructor is an excellent teacher.
- Overall, I learned a great deal from this instructor.

In practice, many universities use global results for summative decisions. This approach has been debated in the research literature. Researchers who are skeptical about global items argue that effective teaching is multifaceted; any attempt to compress all information on what effective teaching entails into a single measure is faulty (for example see Frey, 1978). Teachers favor less global items than specific dimensions of teaching effectiveness (Murray, 1987). However, using global items for summative purposes has been supported by empirical evidence from the literature. For example Abrami, d’Apollonia, and Rosenfield (1997) demonstrated with his factor analysis of 225 items (from 17 student evaluation forms) that one dominant, general factor is common across all student ratings. Abrami (2001), Cashin and Downey (1992) assert that items on specific teaching behaviors are useful for instructors in improving their teaching, while for summative purposes, a shorter form contains global items is able to capture enough information needed for personnel decisions. A more compromising view using weighted scores of measured dimensions measured is promoted by Marsh (2007).

(c) Common items with customization. This approach may well be due to the general understanding that different departments or disciplines have different teaching standards and priorities. Most universities allow departments and instructors to customize items in order to get feedback for teaching improvement. Some universities allow more degrees of customization, for example UBC provides a modular framework with 6 common items for all course evaluations, and faculty members and departments can contribute customized evaluation questions. Many universities provide question banks for such customization so that instructors and departments could select from existing items when constructing an evaluation instrument, such as University of Toronto, Queen’s and University of Michigan.

4.3 PSYCHOMETRIC TESTING

Experts in the field recommend adhering to psychometric principles and practices when developing instruments for effective measurement of teaching performance (for example see Theall & Franklin, 2001). Professionally developed instruments (such as IDEA, SEEQ) usually go through rigorous validation processes involving large sample sizes to ensure sound psychometric qualities. Student evaluation instruments designed by university committees or instructors seldom go through psychometric testing and the relevant revisions (Marsh, 2007; Marsh & Roche, 1997). Researchers in the field proclaim that psychometric analysis is essential even if the final product is an adapted form of a well evaluated instrument, because the latter may not retain the validity of the former in the settings where the instrument is intended to be used (Ory and Ryan, 2001).
4.3.1 VALIDITY

Validity indicates whether or not an instrument measures what it is designed to measure (Wilson, 2005). Validity is the most common concern of student evaluations of teaching and courses in higher education. There is a lack of consensus in the literature regarding validity of student ratings of instruction. Some research has shown that it is more valid than other evaluation tools, although not perfect (such as Cashin, 1988), while others treat student evaluations as only satisfaction or opinion surveys that do not reflect teaching quality.

Despite the evident disputes, there seems to be a trend emerging from the literature regarding validity of student ratings. As noted by Greenwald (1997) “the validity of student rating measures of instructional quality was severely questioned in the 1970s. By the early 1980s, however, most expert opinion viewed student ratings as valid and as worthy of widespread use” (p.1182). Research syntheses continue to show plausible results in demonstrating usefulness of student ratings in determining teaching effectiveness. For example, meta-analyses (d’Apollonia and Abrami, 1996, 1997a, 1997b) showed moderate to high correlations between student ratings of instruction and student achievement (an average of .47). Feldman (1997) also reported correlations from .34-.57 between multiple dimensions of effective teaching and student achievement. And Cohen (1987) reported correlations from .28-.55 across dimensions of effective teaching and achievement, and an average of .43 between achievement and overall rating of the teacher effectiveness (Cohen, 1981) between student achievement and various perspectives of student ratings of teaching effectiveness.15 Researchers are well aware of the limitations of measures of achievement and the complex and imperfect relations between student ratings and their examination performance (Scriven, 1983; Kulik, 2001). Factors other than effective teaching may affect students’ performance in exams, for example as exerted in Kulik (2001), “Bad teachers sometimes put unreasonable pressure on students, and that unethical behavior may produce maximum exam scores.” (p28).

It is important to understand that student achievement, as an important criterion for teaching effectiveness, should not be used as the only condition to determine good teaching (Marsh, 2007). There has been little research examining relations between student evaluation with other aspects of student outcomes such as motivation, study strategies, and career aspirations (Marsh & Roche, 1997).

Nevertheless, researchers are optimistic that student ratings of teaching effectiveness is at least one measure that reflects teaching quality (Centra, 1993). Furthermore, research has provided evidence that student ratings are consistent with other measures, such as trained classroom observers based on instructor behaviors instead of invalid indicators such as personality (Centra, 1993; Murray, 1980), and former student ratings (Howard, Conway & Maxwell, 1985). Instructor self-evaluations also correlate with student ratings (Feldman, 1989; Marsh, Overall & Kesler, 1979) and there is general agreement between students and instructors on the dimensionality of student ratings (Marsh, 2007).

In practice, faculty opinions on student evaluation vary from taking it as reliable, valid and useful to unreliable, invalid and useless (Aleamoni, 1981). Some even consider student ratings as meaningless quantification that leads to personality or popularity contests (Haskell, 1997; Sproule, 2002; Kulik, 2001; Ory, 2001). As a consequence, the use of student ratings may reduce faculty morale (Ryan, Anderson & Birchler, 1980) or encourage undesirable behaviors such as lowering course requirements and become a “teaching teddy bear” (Sacks, 1986), which in turn may compromise student learning. These concerns suggest a disconnection between research and practice as noted by Theall and Franklin (2000). While some concerns may be valid according to researchers, for example, students may not be qualified to judge whether course content is cutting edge or outdated, students are believed to have the ability to

---

15 There has also been research arguing otherwise, for example Rodin & Rodin (1972) reported a negative correlation of .75 between student rating and learning. However such studies are often found to be flawed in methodology (Doyle, 1975; Marsh, 1984).
evaluate aspects such as assignments and interactions between instructor and students in and out of classes (Ory & Ryan, 2001).

The Standards for Educational and Psychological Testing (American Educational Research Association, American Psychological Association, National Council of Measurement in Education) had recently updated the traditional validity framework, which contains content, criterion and construct validity, with a unitary concept that measures the extent to which any evidence supports the intended use and inferences of assessment results (Messick 1989, 1995). According to (Ory & Ryan, 2001, p28), “construct validity should incorporate any evidence that affects the meaning and interpretation of the assessment scores.” Within this framework, validity of student evaluation of teaching is primarily about its inferences on teaching effectiveness and the use and interpretation of the results (2001). There are multiple types of validity evidence, such as content, external, and consequential, etc (2001) which will be discussed more in detail in the final report for those who are interested in the scientific processes of instrument validation. There are also background variables or “noise” that many researchers believe would threaten the validity of student ratings. These variables, or “biases”, have been studied and debated extensively in the literature and will be discussed in a later section in this report. Researchers are more in agreement that the misuse of evaluation results is the major validity threat (for example using the scores to the third decimal points to rank faculty—Theall & Franklin, 2001).

One notable gap in the research literature is how students interpret the evaluation questionnaires, which is an important aspect of validity. For example, do students compare the instructor being evaluated to other instructors when they respond to the evaluation items or do they have a personal standard in mind? And are they well informed with how the evaluation results are going to be used for summative and formative purposes when they respond to the evaluation items. These factors may affect student ratings, for example, research has shown that student ratings are slightly higher when they are told how the results will be used for personnel decisions (Centra, 1976; Feldman, 1979; Overall and Marsh, 1979). Another aspect of validity of student rating resides in the intended and unintended, positive and negative consequences of the interpretation and use of rating scores. Ory & Ryan, (2001) note that this is relatively unexamined territory in the literature and recommend institutions investigate methods and procedures that help enhance intended and positive consequences (such as improving teaching) and reduce unintended and negative ones (such as weakening the curriculum).

4.3.2 RELIABILITY

Reliability is an important aspect of validity. Researchers generally believe that a student evaluation instrument can be reliable (Marsh, 1987), especially when the instrument is well designed and tested according to psychometric principles (Centra, 1993; Aleamoni, 1987; Marsh, 1984). Reliability of a student evaluation tool is usually determined by inter-rater agreement (agreement across different students in the same class) and internal consistency of a scale (correlations among items measuring the same element of effective teaching, such as an instructor’s organization skills, see for example, Abrami, 2001; Theall & Franklin, 2001; Wachtel, 1998; Goldschmid, 1978 ). Research has shown good reliability of teaching evaluations. For example, Feldman (1977) reported that reliability coefficients are usually .80 or higher (with higher coefficients indicating higher reliability). Student ratings also have good internal consistency, that is, the items correlate well with one another. Evaluation scores are also highly correlated across administrations (Ory & Ryan, 2001). There are other aspects of reliability that are not usually examined in teaching evaluations, such as alternative forms measuring the same teaching behaviors (Abrami, 2001).

There has also been abundant evidence demonstrating good stability (rater agreement over time). Longitudinal studies comparing immediate ratings with ratings at least a year after show substantial correlations, for example, in studies by Marsh and Overall (1979) and Overall and Marsh (1980), the average correlation is .83 with 100 courses. In a 13-year longitudinal study by Marsh and Hocevar (1991), no substantial changes of teaching evaluation were found across 195 instructors. Centra (1974)
also reported a correlation of .75 between ratings by students and alumni who had graduated up to 5 years previous.

Generalizability refers to reliable evaluation of an instructor’s general teaching effectiveness rather than one snapshot of his/her performance in a particular course at a particular time. Marsh (1984) investigated this factor by examining 1,364 courses across 4 conditions: a) the same instructor teaching the same courses in two different terms; b) the same instructor teaching two different courses; c) the same course taught by different instructors; and d) two different courses taught by different instructors. Findings indicated that student ratings reflect teaching effectiveness of the instructor rather than the course itself. Researchers recommend that for summative purposes, average scores of several courses taught by the same instructor should be used to enhance generalizability (such as Gilmore, Kane & Naccarato, 1978).

4.4 INSTRUMENT IMPLEMENTATION

Seldin (1993) asserts that poor administration can invalidate a good instrument. Therefore the implementation process is important in ensuring student evaluation will generate useful outcomes. Some common features of current practices in North American universities are:

(a) Student responses are anonymous. Anonymity is an important factor that affects evaluation outcomes. Centra (1979) noted that students who identify themselves tend to give more favorable evaluations, especially if the evaluation is returned to the instructor before final grades. Therefore universities are almost uniformly adopting an anonymous approach and returning feedback to instructors after final grades.

(b) Responses are obtained at the end of the term in the absence of the instructor. Universities usually open student evaluations 2-4 weeks before the final exams to allow adequate time for students to reflect on their evaluations. Research has shown that student ratings are slightly higher when the instructor remains in the room during the evaluation process (Stone, Spool, and Rabinowitz, 1977; Feldman, 1979).

(c) Online vs. paper form. Universities are using different formats of evaluation based on different assumptions of how each type would yield desirable evaluation results. The online formatting provides the ease of response, flexible timing, better reflection and easy data entry. However researchers are concerned that this formatting may not generate the best possible response rate. In the literature, both formats are proven to have their advantages and disadvantages, and no consensus has been reached. Similarly in practice, though most universities are moving their evaluations online (for example in spring 2011, MIT moved to a completely online subject evaluation system), some institutions are doing the opposite. For example, In 2008/09 University of Calgary moved away from an online evaluation process to one that is paper-based and in-class. And only where the paper survey cannot be used (e.g., web-based courses) is an online survey administered.

4.5 USE AND INTERPRETATION OF STUDENT EVALUATION DATA

Student ratings are typically used in the following ways in a postsecondary institution (Ory & Ryan, 2001, p39):

- To help the faculty make teaching and course improvements
- To help the administrators make personnel decisions regarding such things as salary and promotion
- To help the campus select campus teaching award winners
- To help the campus conduct program reviews
- To help the students select courses

The use of the data, especially for personnel decisions is the most contentious aspect of validity of an instrument in the field. Researchers such as McKeachie (1997) suggested that more attention should be directed toward methods of ensuring more valid use. Usually universities norm the global scores as benchmarks (norm-based evaluations), or compare individual instructor’s scores against a pre-determined
standard (criterion-based evaluation), and keep verbal responses confidential to department heads and instructors. Student evaluation is recommended to be used and interpreted in triangulation with other source of information such as teaching portfolios and peer evaluation.

Researchers express that despite the concerns and research findings, in practice teaching evaluations are primarily used for summative personnel decisions (for example see Beran, Violato, & Kline, 2007; Wagenaar, 1995). Although theoretically student evaluations provide valuable feedback on teaching behaviors, research found that student evaluation results are seldom used by faculty in improving their teaching quality.

Abrami (2001) recommended strategies to improve use of student rating data, such as statistical and administration strategies, and set criteria/standards in advance. For example whether to use criteria- or norm-referenced approaches, averaging global items and scores for courses combined, weighing, and data presentation format such as graphs.
5 POTENTIAL BIASES: MYTHS OR REALITY?

A plethora of research has been focusing on variables that could potentially threaten validity of student ratings by introducing biases to the evaluation outcome (as described by Abrami, 2001, Marsh, 2007, Marsh & Roche, 1997; Theall & Franklin, 2001). Researchers have been debating on whether or not these variables could in fact bias student ratings or their relation to evaluation scores are simply evidence for construct validity (Marsh, 2007; Centra, 2003)\(^16\). Much research has suggested that most of these variables are not in fact conclusive biases, but myths (Aleamoni, 1999) at most. Marsh (2007) uses a more neutral term “background characteristics” before concluding whether or not a variable introduces bias to student evaluations, such as class size, workload/difficulty, and grading leniency. He pointed out that most bias research is flawed in methodology and unclear with defining what constitutes bias\(^17\).

Perceptions of potential biases of student evaluation of teaching have been a source of anxiety for instructors (Franklin & Theall, 1989). Such perceptions are based on assumptions such as students are not qualified to evaluate teaching (Nasser & Fresko, 2002); student evaluation is in fact a popularity contest where individuals who are warm, charismatic and give students higher grades are always winners (Aleamoni, 1999; Baldwin & Blattner, 2003; Williams & Ceci, 1997). Although there has been abundant research refuting such assumptions, in practice anxiety and concerns remain (Aleamoni, 1999; Ory, 2001; Theall & Feldman, 2007).

Marsh (2007, p348-349) provided a list of most common concerns from the literature and their effects on student evaluations. Other than the apparent methodological flaws, most effects found on these variables are either small, inconsistent, or inconclusive\(^18\). More importantly, it is argued that most of these variables do not constitute bias toward the validity of student ratings\(^19\).

5.1 PRIOR SUBJECT INTEREST

This variable has been found to correlate with certain aspects of effective teaching, such as learning/value (correlation sizes approximately .4). Thus it is argued to be a source of validity instead of a bias. And the same effects hold across measures of teaching effectiveness, such as instructor self-evaluation.

5.2 EXPECTED GRADE/GIVEN GRADES

Given the widespread concerns of this specific effect, many approaches have been taken to investigate this potential bias in the literature. Correlations between this variable and student ratings tend to be small (.11-.20, Marsh, 2007) or inconsistent. It is difficult to separate the effects between grading leniency and student received grades, which may reflect student learning. Measures of grading leniency are also under question. Additionally, research that reports positive correlations between grades and ratings, for example Greenwald and Gilmore, (1997), often neglects how grading standards vary across subjects, disciplines, and course levels, and consequently deems the bias claim non-generalizable (Kulik,

---

\(^{16}\) This will be discussed in more details in the Final Report after data collection and analysis for the Project.

\(^{17}\) In fact some concerns have been proven to be based on poorly designed research or irrelevant to student rating, discussions and debates on such topics are now deemed passé (Gravestock & Gregor-Greenleaf, 2008), such as the “Dr. Fox effect”: In study by Naftulin, ware, and Donnelly (1973), a professional actor was hired to deliver a lecture on mathematical game theory to medical educators posing as an expert in the field. His lecture was rated favorably by the audience even though it involves mainly incorrect information. This effect caused by showmanship is termed as the “Dr. Fox Effect”.

\(^{18}\) An in-depth discussion of potential biases is beyond the scope of this report. A few variables are briefly discussed here based on their relation to student ratings as examples. Please see Marsh (2007) and Centra, (2003) for closer examinations of the variables. These variables will be taken into consideration when developing and validating our instruments. More thorough discussions will be provided in the Final Report for those who are interested in the validation process of our instrument.

\(^{19}\) Marsh (2007) argues that: “If a potential biasing factor actually does have a valid influence on teaching effectiveness and this influence is evident in different indicators of teaching effectiveness… then it may be possible that the influence reflects support for the validity of SETs(student evaluation of teaching), rather than a bias (p349).”
All the research suggests that this variable may generate only weak, if any, biasing effects toward student ratings.

### 5.3 Workload/Difficulty

In contrast to common concerns that instructors who assign more work and give difficult assignments tend to receive less favorable ratings, some research found that there is a positive correlation between workload/difficulty and student ratings. Instructors’ self-evaluation shows consistent effects with this variable. Although further analyses show that the relationship of this variable and student ratings can be more complicated and non-linear, Marsh (2007) argued that it is not a biasing factor toward student ratings.

### 5.4 Class Size

This variable is often found to have a small negative correlation with student ratings and moderate effects on dimensions such as group interactions and individual rapport. Logically a smaller class allows the instructor to allocate more time and attention to each student. The relationship between class size and student ratings tends to be non-linear, with very large classes receiving similar student ratings as small classes. Caution is needed when interpreting such effects. For example favorable student ratings in very large classes may reflect appropriate teaching strategies in such settings. Additionally instructors’ self-evaluation tends to be consistent with student ratings on specific aspects of effective teaching, suggesting that this variable is a source of evidence for validity rather than a bias.

In order to alleviate possible anxieties from faculty it is important to acknowledge to the SFU community that the project team is aware of various sources of possible biases and factors that may threaten validity when developing our evaluation instruments. These can be controlled, for example, through data triangulation and appropriately controlling for differences in student ratings associated with these background characteristics when interpreting evaluation results.

Bear in mind that even if all the above concerns were addressed by research results, experts caution that “…data can be and are misused on a regular basis. Even if ratings results were perfectly reliable and valid (and no educational, psychological, or sociological instrument provides data that are perfect), misuse would still be a major problem.” (Theall & Franklin, 2001, p46). It is important to inform the SFU community that our project is not striving for perfection, but is developing a valid and reliable instrument that best suits SFU’s evaluation needs by closely examining the relevant research literature, as well as through a thorough community engagement during the process of instrument development and implementation.
6 REFERENCES


